Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220460

Building an i2b2-Based Population Repository for COVID-19 Research

Miguel PEDRERA-JIMENEZ[a,b,1], Noelia GARCIA-BARRIO[a], Gema HERNANDEZ-IBARBURU[c], Blanca BASELGA[a], Alvar BLANCO[a], Fernando CALVO-BOYERO[a], Alba GUTIERREZ-SACRISTAN[d], Víctor QUIROS[a], Juan Luis CRUZ-BERMUDEZ[a], José Luis BERNAL[a], Laura MELONI[e], David PEREZ-REY[c], Matvey PALCHUK[d,e], Isaac KOHANE[d] and Pablo SERRANO[a]

[a]*Data Science Unit, Research Institute Hospital 12 de Octubre, Madrid, Spain*
[b]*ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain*
[c]*Biomedical Informatics Group, Universidad Politécnica de Madrid, Madrid, Spain*
[d]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*
[e]*TriNetX, LLC, Cambridge, MA, USA*

Abstract. Reuse of Electronic Health Records (EHRs) for specific diseases such as COVID-19 requires data to be recorded and persisted according to international standards. Since the beginning of the COVID-19 pandemic, Hospital Universitario 12 de Octubre (H12O) evolved its EHRs: it identified, modeled and standardized the concepts related to this new disease in an agile, flexible and staged way. Thus, data from more than 200,000 COVID-19 cases were extracted, transformed, and loaded into an i2b2 repository. This effort allowed H12O to share data with worldwide networks such as the TriNetX platform and the 4CE Consortium.

Keywords. Electronic Health Records, Real World Data, Data Reusability, Semantics, Standardized repositories, i2b2, TriNetX, 4CE Consortium, COVID-19.

1. Introduction

COVID-19 pandemic has been the major health challenge in recent decades, being declared a pandemic state on March 11, 2020 by World Health Organization (WHO) [1]. In this critical situation, Electronic Health Records (EHRs) have proven crucial for patient management and healthcare, as well as for clinical research on this new and unknown disease [2].

In this sense, the reuse of EHRs for COVID-19 research required health data to be available in an agile way, flexible to specific requirements, standardized according to international recommendations and exploitable through advanced analysis tools [3]. Some of the health data resources that facilitate this task are the standardized clinical repositories such as i2b2 and OMOP CDM [4, 5].

Thus, this work aims to describe how Hospital Universitario 12 de Octubre in Madrid, Spain (H12O) has developed and used its i2b2 repository [6] to respond to the data needs that arose during the COVID-19 pandemic.

[1] Corresponding Author, Miguel Pedrera Jiménez, Data Science Group, Hospital Universitario 12 de Octubre, Av. de Andalucía S/N, 28041, Madrid, España; E-mail: miguel.pedrera@salud.madrid.org.

2. Methods

This work was carried out at H12O, implementing this organization's methodology to effectively reuse EHRs for research [3, 7]. The sources and type of data required, standardization resources, mapping efforts, and data volumes obtained are provided in this paper to help other healthcare institutions willing to share COVID-19 data.

2.1. EHRs from Hospital Universitario 12 de Octubre

The source for data extraction was the healthcare information systems of H12O, which have been formally modeled through health information standards such as ISO 13606, and incorporate terminologies such as SNOMED CT and LOINC. This standardization of the EHRs has allowed their reuse, without additional manual efforts and full meaning, in data collection processes for research and other secondary uses [3].

To make EHRs reuse efficient, we designed a set of multipurpose information models, common to all use cases. These were formalized through clinical archetypes, and implemented in the healthcare information systems of the hospital. Hence, these standardized information resources have been used to implement research tools at H12O, including a clinical repository based on the i2b2 data model [6]. Table 1 describes this set of information models for EHRs reuse.

Table 1. Information models for EHRs reuse of Hospital Universitario 12 de Octubre.

Archetype	Description	Terminology binding
Patient	Demographics data, e.g., birthdate, sex and vital status.	SNOMED CT
Encounter	Data related to inpatient, emergency and outpatient visits.	SNOMED CT
Location	Patient locations during hospitalization, e.g., ICU admission.	SNOMED CT
Observation	Clinical, laboratory and patient-reported observations.	SNOMED CT, LOINC
Diagnosis	Health issues and clinical diagnoses.	SNOMED CT
Medication	Pharmacological treatment prescribed.	SNOMED CT
Procedure	Procedures performed, e.g., surgeries and nursing interventions.	SNOMED CT

2.2. Informatics for Integrating Biology & the Bedside (i2b2)

The i2b2 tool is a scalable informatics framework that organizes and transforms patient-oriented health data in a way that is optimized for research. It was developed by Harvard Medical School with funding from the National Institutes of Health (NIH).

The database design for the clinical repository is based on a star schema composed of a central table for the observed clinical facts, which is related to five additional tables that provide contextual information about the observation, i.e., patient, visit, provider, and observable entity. Figure 1 shows the data model proposed by i2b2.

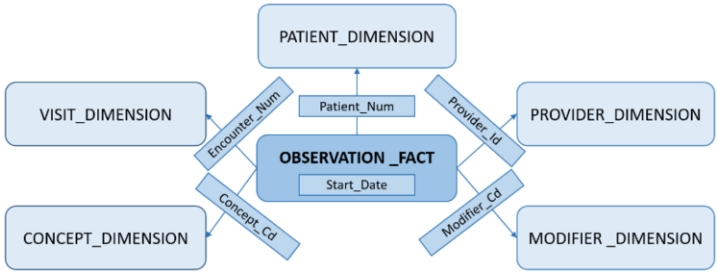

Figure 1. Data model for clinical repository proposed by i2b2.

2.3. COVID-19-related concepts

COVID-19 meant facing a new and unknown disease, so healthcare information systems had to be evolved, in an agile, flexible and staged way, to include these new concepts necessary for diagnosis, treatment and prevention of this condition [2].

This set of concepts was identified through a group of clinical domain experts at H12O. For this purpose, we adopted the WHO recommendations for patient stratification during the pandemic [8], and the Spanish Ministry of Health terminology standards [9]. Table 2 shows the set of COVID-19-related concepts that were modeled and standardized, indicating the terminology, and the date when these concepts were implemented into healthcare information systems.

Table 2. Standardized set of COVID-19-related concepts of H12O ordered by implementation date.

EHRs concept	Description	Since
SNOMED-CT: 63681000122103	Diagnoses disease caused by 2019 novel coronavirus with specific diagnostic tests	2020-03-08
SNOMED-CT: 840544004	Suspected disease caused by 2019 novel coronavirus	2020-03-08
SNOMED-CT: 63341000122104	Type of respiratory support	2020-03-08
SNOMED-CT: 704296008	At risk of impaired respiratory system function	2020-03-08
LOINC: 94315-9	SARS-related coronavirus E gene [Presence] in Specimen by NAA with probe detection	2020-03-17
SNOMED-CT: 62811000122102	Diagnoses disease caused by 2019 novel coronavirus without specific diagnostic tests	2020-03-17
SNOMED-CT: 160734000	Lives in a nursing home	2020-03-20
SNOMED-CT: 688232241000119100	Discarded disease caused by 2019 novel coronavirus	2020-04-15
SNOMED-CT: 736534008	EuroQol five dimension five level index value (observable entity)|	2020-06-29
LOINC: 94558-4	SARS-CoV-2 (COVID-19) Ag [Presence] in Respiratory specimen by Rapid immunoassay	2020-09-29
LOINC: 96751-3	SARS-CoV-2 (COVID-19) S gene mutation detected [Identifier] in Specimen by Molecular genetics method	2020-12-23
LOINC: 96895-8	SARS-CoV-2 (COVID-19) lineage [Identifier] in Specimen by Molecular genetics method	2020-12-23
SNOMED-CT: 1156257007	Administration of vaccine product against Severe acute respiratory syndrome coronavirus 2	2021-01-10

3. Results

Our main results are to put in place a whole extraction, transformation, and loading process (ETL) of the patient-level COVID-19 related data into an i2b2 repository and the successful usage for research across multiple projects.

3.1. Extraction, transformation and loading of COVID-19 data

First, the data was extracted, transformed (operation T.1.2, change of coding system [7]), and loaded into the i2b2 repository. Table 3 shows, for each concept related to COVID-19, the terminology mapping implemented, the volume of records and the number of patients loaded on January 17, 2022.

Table 3. Volume of records and patients related to COVID-19 loaded into the i2b2 repository.

EHRs concept	Mapped concepts	Records (N)	Patients (N)
SNOMED-CT: 63681000122103	ICD10CM:U07.1	56,051	53,300
SNOMED-CT: 840544004	ICD10CM:Z20.822	9697	8415
SNOMED-CT: 63341000122104	ICD10PCS:5A09, ICD10PCS:5A19	8274	7835
SNOMED-CT: 704296008	*No mapping required*	7427	7026
LOINC: 94315-9	LOINC: 94315-9	355,432	211,656
SNOMED-CT: 62811000122102	ICD10CM:U07.2	1402	1307
SNOMED-CT: 160734000	ICD10CM:Y92.12	2277	2240
SNOMED-CT: 688232241000119100	ICD10CM: Z03. 818	24,386	22,084
SNOMED-CT: 736534008	LOINC: 97332-1	3250	1184
LOINC: 94558-4	LOINC: 94558-4	46,850	35,441
LOINC: 96751-3	LOINC: 96751-3	1124	1094
LOINC: 96895-8	LOINC: 96895-8	13,906	13,152
SNOMED-CT:1156257007	RXNORM:2468231	279,450	160,738

3.2. Use of COVID-19 data for research

COVD-19 data stored into the H12O i2b2 repository have been used in several projects, including two international ones, the TriNetX platform and the 4CE Consortium.

3.2.1. TriNetX platform

The TriNetX platform is a global health research network to share real-world data, making clinical and observational research more accessible and efficient [10]. This platform combines real-time access to longitudinal clinical data with state-of-the-art analytics to optimize protocol design and feasibility, site selection, patient recruitment, and enable discoveries through the generation of real-world evidence. It contains data from more than 120 sites in 19 countries.

Data on more than 200,000 COVID-19 cases (suspected, diagnosed, confirmed and discarded) were loaded in real-time as of March 8, 2020. This allowed us to respond to questions of general interest from the beginning of the pandemic. One of the most important was to determine the impact on the patient outcome of coming to the hospital from a nursing home [11]. Similarly, this tool has allowed the development of complex clinical studies of high impact, combining COVID-19 data with data from other conditions. Among other initiatives, a relevant study was developed by H12O Hematology Department, which confirmed that the COVID-19 pandemic has a more severe impact on patients with Multiple Myeloma (MM) than non-MM patients [12].

3.2.2. 4CE Consortium

The 4CE Consortium federates more than 300 hospitals from seven countries, and so far, the H12O is the only Spanish hospital [13]. This was possible thanks to the joint efforts from the i2b2 community and the health data experts from H12O. Thus, having the data in the i2b2 repository and having as reference the extraction and transformation scripts developed by the 4CE Consortium, we could obtain the required data in a timely manner.

Hence, we could extract and aggregate patient-level data from 7,028 COVID-19 cases. This work, and the international effort of this consortium, allowed doing relevant research on patients affected by COVID-19. In one of these, data from 671 children hospitalized were included in a pediatric study, being 78 of them from H12O [14].

4. Conclusions

The rapid identification, modeling and standardization of COVID-19 concepts into EHRs allowed obtaining valuable data for research on this new disease. These data were loaded in an agile, flexible and staged manner into the i2b2 repository of H12O for having data available according to standards and exploitable by advanced analysis tools.

Thus, the TriNetX platform was used to answer research questions since the beginning of the COVID-19 pandemic. Similarly, H12O joined the 4CE Consortium, combining data with other hospitals, and participating in relevant international studies.

Acknowledgment

This work has been supported by Research Projects PI18/00981, PI18/01047 and PI18CIII/00019; funded by Instituto de Salud Carlos III, co-funded by ERDF/ESF.

We want to thank TriNetX and 4CE Consortium for their support in this project.

References

[1] Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;382:727–33. doi:10.1056/NEJMoa2001017.

[2] Pedrera M, Garcia N, Blanco A, et al. Use of EHRs in a Tertiary Hospital During COVID-19 Pandemic: A Multi-Purpose Approach Based on Standards. Stud Health Technol Inform. 2021;281:28-32. doi:10.3233/SHTI210114.

[3] Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, et al. Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. J Biomed Inform. 2021;115:103697. doi:10.1016/j.jbi.2021.103697.

[4] Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17(2):124-130. doi:10.1136/jamia.2009.000893.

[5] Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-578.

[6] González L, Pérez-Rey D, Alonso E, et al. Building an I2B2-Based Population Repository for Clinical Research. *Stud Health Technol Inform*. 2020;270:78-82. doi:10.3233/SHTI200126.

[7] Pedrera M, Garcia N, Rubio P, Cruz JL, Bernal JL, Serrano P. Making EHRs Reusable: A Common Framework of Data Operations. Stud Health Technol Inform. 2021;287:129-133. doi:10.3233/SHTI210831.

[8] WHO ICD-10 codes for COVID-19. https://www.who.int/classifications/icd/COVID-19-coding-icd10.pdf. Accessed January 17. 2021.

[9] Health Ministry of Spain: Minimum data set for clinical reports. https://www.boe.es/eli/es/rd/2010/09/03/1093. Accessed January 17. 2021.

[10] TriNetX Platform. https://trinetx.com/. Accessed January 17. 2021.

[11] COVID-19 and nursing homes at H12O. https://www.comunidad.madrid/noticias/2021/12/16/hospital-12-octubre-premio-barea-su-trabajo-centros-socio-sanitarios. Accessed January 17. 2021.

[12] Martinez-Lopez J, Hernandez-Ibarburu G, Alonso R, et al. Impact of COVID-19 in patients with multiple myeloma based on a global data network. Blood Cancer J. 2021;11(12):198. Published 2021 Dec 10. doi:10.1038/s41408-021-00588-z.

[13] Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020;3:109. Published 2020 Aug 19. doi:10.1038/s41746-020-00308-0.

[14] Bourgeois FT, Gutiérrez-Sacristán A, Keller MS, et al. International Analysis of Electronic Health Records of Children and Youth Hospitalized With COVID-19 Infection in 6 Countries [published correction appears in JAMA Netw Open. 2021 Jul 1;4(7):e2122388]. JAMA Netw Open. 2021;4(6):e2112596. Published 2021 Jun 1. doi:10.1001/jamanetworkopen.2021.12596.